Claims

What is claimed is:

1. A method of detecting one or more outliers in a data set, comprising the steps of:

determining one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and

determining one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

2. The method of claim 1, wherein a range is defined as a set of contiguous values on a given dimension.

3. The method of claim 1, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

4. The method of claim 3, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1-f^k)}}$, where k represents the number of dimensions in the data set, f represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

5. The method of claim 3, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

6. The method of claim 1, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

7. The method of claim 6, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

8. The method of claim 6, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

9. The method of claim 6, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

10. A method of detecting one or more outliers in a data set, comprising the steps of:

identifying and mining one or more patterns in the data set which have abnormally low presence not due to randomness; and

identifying one or more records which have the one or more patterns present in them as the one or more outliers.

11. Apparatus for detecting one or more outliers in a data set, comprising:

at least one processor operative to: (i) determine one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and (ii) determine one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

12. The apparatus of claim 11, wherein a range is defined as a set of contiguous values on a given dimension.

13. The apparatus of claim 11, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

14. The apparatus of claim 13, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1-f^k)}}$, where k represents the number of dimensions in the data set, f represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

15. The apparatus of claim 13, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

16. The apparatus of claim 11, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

17. The apparatus of claim 16, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

18. The apparatus of claim 16, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

19. The apparatus of claim 16, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

20. Apparatus for detecting one or more outliers in a data set, comprising:

at least one processor operative to: (i) identify and mine one or more patterns in the data set which have abnormally low presence not due to randomness; and (ii) identify one or more records which have the one or more patterns present in them as the one or more outliers.

21. An article of manufacture for detecting one or more outliers in a data set, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

determining one or more sets of dimensions and corresponding ranges in the data set which are sparse in density; and

determining one or more data points in the data set which contain these sets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

22. The article of claim 21, wherein a range is defined as a set of contiguous values on a given dimension.

23. The article of claim 21, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

24. The article of claim 23, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1-f^k)}}$, where k represents the number of dimensions in the data set, f

represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

25. The article of claim 23, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

26. The article of claim 21, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

27. The article of claim 26, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

28. The article of claim 26, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

29. The article of claim 26, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

30. An article of manufacture for detecting one or more outliers in a data set, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

identifying and mining one or more patterns in the data set which have abnormally low presence not due to randomness; and

identifying one or more records which have the one or more patterns present in them as the one or more outliers.